面向临床决策的电子病历文本潜在语义分析*

李国垒[1]　陈先来[1,2,3]　夏　冬[4]　杨　荣[5]
[1](中南大学信息安全与大数据研究院　长沙 410013)
[2](医学信息研究湖南省普通高等学校重点实验室(中南大学)　长沙 410013)
[3](湖南省高等学校医学大数据 2011 协同创新中心　长沙 410013)
[4](中国科学院成都文献情报中心　成都 610041)
[5](中南大学湘雅医院　长沙 410078)

摘要:【目的】通过对电子病历中重要文本进行语义分析，提取辅助临床治疗方案选择的决策知识，实现电子病历的临床决策支持功能。【方法】使用词典和统计相结合的分词算法，对训练样本中出院记录文本进行分词处理，从中提取临床术语及治疗方案，并对其进行潜在语义分析，找出临床术语与治疗方案之间的潜在语义联系，建立胃癌治疗方案辅助选择的潜在语义模型。【结果】利用测试样本对语义模型进行测试，在三维语义空间内，发现 1 000 份测试样本中有 605 份可以从临床症状的描述准确地推算出其所对应的治疗方案，正确率为 60.5%。【局限】仅以出院记录文本为研究对象，没有对其他病历文本进行分词处理。【结论】潜在语义分析方法能够有效地处理临床文本，辅助医生的临床决策，对于电子病历的开发应用具有重要意义。

关键词: 电子病历　中文文本切分　潜在语义分析　胃癌　临床决策支持　治疗方案选择
分类号: G350

1　引　言

病历是一种十分重要的医学信息资源。我国卫生信息化进程的加速使得电子病历逐渐普及，而电子病历的价值也越来越受到相关领域专家学者的重视。电子病历中除了结构化数据，还存在大量非结构化数据，既有规范的临床术语，也有不规范的自然语言。从电子病历中提取知识用于临床决策，已成为近年来电子病历应用研究中亟待解决的主要问题之一。

全球胃癌的发病率和死亡率居恶性肿瘤的第 2 位和第 3 位，胃癌的诊疗是人们高度关注的研究热点。对每一个胃癌病人而言，合理的个性化治疗方案是提高治愈率和取得良好治疗效果的前提。对于采取何种治疗方案，目前主要还是依靠医生的主观经验。随着电子病历的普及，依据大规模历史病历文本建立临床决策支持系统，辅助医务人员开展临床诊疗工作，对提高胃癌的治疗效果，提高电子病历的使用价值，具有十分重要的意义。

2　研究背景

目前，国内在电子病历结构化方面，由于中文语言的特殊性，医护人员通常以自由文本的形式对患者的相关临床信息进行描述。医护人员在使用这些描述性语言时，用词不受约束，可以使用任意词汇、代码或者缩略词。这种方式会造成临床数据共享困难，不利于临床数据的利用和临床医疗决策支持系统的应用。因此国内对于电子病历系统的研究，还是侧重于电子病历系统编辑器的研究，并没有实

通讯作者：杨荣，ORCID：0000-0003-1433-9606，E-mail：cxlyr0576@163.com。
*本文系国家社会科学基金项目“面向临床决策的电子病历潜在语义分析及应用研究”(项目编号:13BTQ052)的研究成果之一。

现具有较强临床决策支持功能的电子病历系统。从电子病历中提取知识，就要使病历内容能被计算机“理解”，而语义分析则是计算机“理解”病历内容的有效方法。

潜在语义分析(Latent Semantic Analysis, LSA)作为一种用于知识获取、归纳和展示的计算理论和方法，具有可计算性强、主观影响因素少等优点[1]。它利用统计学方法对文本集进行处理，从而提取出词的潜在语义结构，这种潜在语义结构即是词语在上下文语境信息的总和。经过 20 多年的发展，LSA 理论与技术已较为完善、成熟，且已被用于处理医学信息，Cohen 等利用潜在语义分析法构建了精神病术语的语义空间[2]，并探索抽取相关概念之间的语义关系[3]。Ginter 等应用潜在语义分析法和隐马尔可夫模型对分割、标识护理文档主题开展研究[4]。Wild 等结合网络分析和潜在语义分析法对医学概念的发展进行研究[5]。Wang 等依据病历信息，利用 LSA 实现了生物医学时间序列的自动聚类[6]，Abate 等则依托 PubMed 文献数据库，利用潜在语义分析法将生物医学文献中的生物医学术语的相关关系进行量化[7]。国内甘艳芳等利用 LSA 计算中医证候间的相关关系[8]；雷蕾等基于中医药文献数据，使用概率潜在语义分析算法研究中药配伍方案，为中药处方发现提供新途径[9]。目前国内主要针对已有的较为规范的临床术语或医学文献主题词语义关系进行研究，还没有对于临床实践应用较多的大量的非规范化临床用语进行相关语义分析的研究。

本文收集了 1 500 份胃肿瘤病例的出院记录，抽取其中的文本作为研究对象。以中国科学院计算技术研究所 ICTCLAS 分词系统为基础，对出院记录文本的词语切分进行探索，利用中国生物医学文献数据库术语、基于互信息的统计方法对出院记录文本进行分词处理，从中抽取临床术语；制定胃肿瘤治疗方案自动抽取规则，并使用 Python 编写脚本从出院记录中提取了胃肿瘤治疗方案；利用潜在语义分析方法建立胃肿瘤治疗方案选择决策支持模型，并对其进行评价。在此过程中，构建了临床术语与治疗方案的共现矩阵，并利用 NumPy 进行矩阵的奇异值分解，得到临床术语和治疗方案在语义空间中的坐标向量，计算出两者之间的相关度，按照相关度大小抽取决策规则，初步完成了胃肿瘤治疗方案选择支持模型的建立，并利用另外的 1 000 份病例记录作为测试样本，完成模型的验证，为临床决策支持系统的开发奠定基础。

3　研究对象与方法

3.1　数据来源

从湖南省多所三甲医院中提取 2 500 份 2010 年–2014 年间第一诊断为胃癌(ICD-10 编码: C16, D00.2)的病历，随机抽取 1 500 份用于训练，其余 1 000 份用于测试。

病历中的出院记录主要包含住院病情摘要、诊治经过、出院时情况和出院医嘱等模块，是患者完整电子病历的高度浓缩。其中，住院病情摘要模块详细记录了患者住院时的临床症状及检查检验结果，诊治经过模块中详细记录了患者的整个诊治过程。患者出院记录部分内容示例如图 1 所示:

中年男性，腹胀半年。体查：体温 36.5℃ 脉搏 80 次/分 呼吸 20 次/分 血压 125/70mmHg 皮肤粘膜色泽正常，巩膜无黄染。肺部无异常，心前区无隆起，心尖搏动正常，位于左侧第 5 肋间锁骨中线内侧 0.5cm。肺部无异常，全腹柔软，右下腹可触及一肿块，无压痛，无反跳痛；肝脏剑突下未及，肋下未及；脾脏肋下未及；肾脏未及。肝区无叩痛，双肾区无叩痛，移动性浊音阳性。肠鸣音 3 次/分，音正常。脊柱正常，棘突无压痛，无叩痛，四肢活动正常，双侧下肢无水肿。肛门外生殖器正常。生理反射正常，肌张力正常，肌力 5 级。入院后完善相关检查，血常规、肝肾功能电解质、心电图基本正常。胸片示双上肺陈旧性结核，右上肺结核球形，建议进一步 CT 检查。ECG 示窦性心动过缓;T 波改变。腹部 CT 示上腹部网膜增厚并多发淋巴结肿大右肝实质性结节腹水。考虑胃癌腹腔广泛转移。于 2013-06-19 予奥沙利铂针+替吉奥胶囊化疗，辅以护胃护肝护心、止呕等对症支持治疗。

图 1　出院记录部分内容示例

3.2　治疗方案的抽取

通过咨询相关临床医生并参考《胃癌规范化诊疗指南(试行)》[10]，将胃癌的治疗方案归纳为手术治疗、放化疗、手术治疗与放化疗相结合以及对症治疗 4 种。在翻阅大量病历资料和咨询相关医生的基础上，发现出院记录中若采取手术治疗，通常会有与手术相关的字样；如采取放化疗方案，也会有与放化疗相关的字样。通过对病历信息的查阅，得到治疗方案与相关字样的对应关系如表 1 所示:

表 1　治疗方案与相关字样对应关系

治疗方案	相关字样
手术治疗	全麻、(全)麻、行胃镜下、胃癌根治术、胃大部分切除术、局麻、全胃切除等
放化疗	化疗、放疗和放化疗相关药品名称

根据病历中“诊治经过”的文本内容，制定如下规则，用于抽取治疗方案:

(1) 若其中含有“手术”相关字样而不含“放化疗”相关字样，则治疗方案被认定为手术治疗;

(2) 若其中含有“手术”相关字样且含有“放化疗”相关字样，则治疗方案被认定为手术+放化疗;

(3) 若其中含有“放化疗”相关字样，且不含“手术”相关字样，则治疗方案被认定为放化疗;

(4) 若其中既无“手术”相关字样也无“放化疗”相关字样，则治疗方案被认定为对症治疗。

按照上述规则，利用 Python 编写程序，对 2 500 份出院记录(包括 1 500 份训练样本和 1 000 份测试样本)进行处理，提取出每一份出院记录中的治疗方案。结果发现，其中 1 102 例采取手术治疗，286 例采取手术+放化疗，457 例采取放化疗，655 例对症治疗。例如，对图 1 所示的出院记录进行治疗方案抽取，结果为放化疗，这与实际结果是一致的。从 2 500 份样本中随机抽取 100 份出院记录，进行人工核查，发现有 95 份抽取出的治疗方案与实际相符合，正确率达到 95%，证实该抽取方案是可行的。

3.3 临床术语的抽取

本文利用自定义词典结合统计分词的方法抽取病历文本中的临床术语。实验在 MyEclipse 集成开发环境下，参考 ICTCLAS 5.0 分词系统提供的 API，采用 Java 语言实现对出院记录文本的分词处理[11]。具体步骤如图 2 所示:

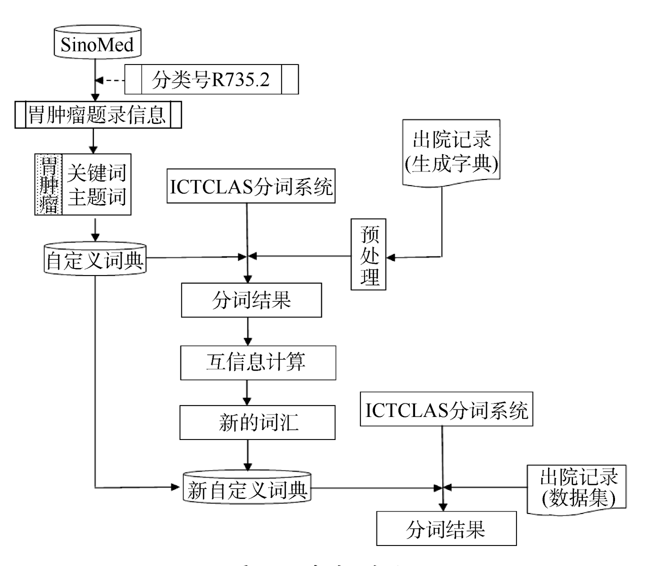


图 2 分词流程

以中国生物医学文献数据库(SinoMed)为依据，构建自定义词典。SinoMed 中包含大量的临床研究文献，并且在题录中有规范的主题词和关键词。本研究以分类号“R735.2”检索 SinoMed 中 2001 年–2003 年的文献(近几年的文献题录中只标注了关键词，没有标注主题词)，共检索到有效文献题录 4 244 条，抽取其中的关键词和主题词，合并去重后得到自定义词典，共含有 5 429 个词汇。

采用统计方法进一步完善自定义词典。利用 ICTCLAS 分词系统，使用以上步骤所构建的自定义词典，对 1 500 份训练样本数据进行词语切分后，直接对切分结果进行统计分析，计算相邻词汇的互信息值[12]，公式为:

$$MI(A,B)=\log_2\frac{P(A,B)}{P(A)P(B)} \tag{1}$$

其中，P(A, B)表示词汇 A, B 在所有记录中共现频次的概率，P(A)表示出现词汇 A 的概率，P(B)表示出现词汇 B 的概率。在实验文本中，相邻两个字组合生成新的词汇共有 51 658 种可能，其中互信息值大于等于 0 的词对共有 11 845 种，将其与之前构建的自定义词典中的 5 429 个词汇进行合并去重，构成新的自定义词典，共含有词汇 17 113 个。

利用所建自定义词典对医疗文本进行切分。依据所构建的自定义词典，再次对 1 500 份训练样本出院记录进行切分。图 1 的出院记录切分的结果如图 3 所示，发现词典分词结合统计分词的方法可以弥补 ICTCLAS 系统的不足，能够将 ICTCLAS 系统拆分开的词汇有效地合并起来。例如，可以将“无黄染”、“无压痛”、“外生殖器”等临床症状描述的术语准确地切分出来。

中年 男性 腹胀 半 年 。 体查 ： 体温 36.5 ℃ 脉搏 80 次/分 呼吸 20 次/分 血压 125 / 70mmHg 皮肤 粘膜色泽 正常 ， 巩膜 无黄染 。 肺部 无异常 ， 心前区 无隆起 ， 心尖 搏动 正常 ， 位于 左侧 第5 肋间 锁骨中线 内侧 0.5cm 。 肺部 无异常 ， 全腹 柔软 ， 右下腹 可 触及 一 肿块 ， 无压痛 ， 无反跳痛 ； 肝脏 剑突 下 未及 ， 肋下 未及 ； 脾脏 肋下 未及 ； 肾脏 未及 。 肝区 无叩痛 ， 双肾 区 无叩痛 ， 移动性 浊音 阳性 。 肠鸣音 3 次/分 ， 音 正常 。 脊柱 正常 ， 棘突 无压痛 ， 无叩痛 ， 四肢 动 正常 ， 双侧 下肢 无水肿 。 肛门 外生殖器 正常 。 生理反射 正常 ， 肌张力 正常 ， 肌力 5级 。 入院 后 完善 相关 检查 ， 血常规 、 肝肾 功能 电解质 、 心电图 基本 正常 。 胸片 示 双上肺 陈旧 性 结核 ， 右上肺 结核 球形 ， 建议 进一步 CT检查 。 ECG 示 窦性心动过缓 ; T波 改变 。 腹部 CT 示 上腹部 网膜 增厚 并 多发 淋巴结 肿大 右肝 实质性 结节 腹水 。 考虑 胃癌 腹腔 广泛转移 。 于 2013-06-19 予 奥沙利铂 针 + 替吉奥 胶囊 化疗 ， 辅以 护胃 护肝 护心 、 止呕 等 对症支持 治疗

图 3 分词结果示例

采用 ICTCLAS+自定义词典+统计分词策略对 1 500 份训练样本进行切分，并统计每一个切分出来的临床术语的出现频次，剔除掉标点符号、数字、日期、单个字符等特殊字符以及与治疗方案相关的手术名称与放化疗药品名称等，结果如表 2 所示：

表 2 临床术语频次统计

序号	词汇	频次
1	未见	3 864
2	正常	3 346
3	入院	3 239
4	患者	3 031
5	未触及	2 662
6	淋巴结	2 472
……	……	……
1091	表面凹凸不平	15
1092	两端	15
……	……	……
7209	唐*秋	1
7210	段*勤	1

为了防止临床术语与治疗方案共现频次矩阵过于稀疏而影响计算结果，本实验只抽取临床术语出现频次大于等于 15 的词汇，得到词汇 1 092 条。

3.4 潜在语义空间的构造

(1) 临床术语–治疗方案矩阵的构建

在提取临床术语、治疗方案的基础上，根据临床术语与治疗方案之间的共现情况，构建临床术语–治疗方案矩阵。利用 Java 编制程序，统计 1 500 份训练样本中临床术语与治疗方案在出院记录中的共现频次，生成临床术语–治疗方案矩阵 H(1092 × 4)，其部分示例如表 3 所示：

表 3 临床术语–治疗方案共现矩阵 H(部分示例)

临床症状	治疗方案			
	手术+放化疗	手术	对症治疗	放化疗
…	…	…	…	…
无压痛	116	308	168	127
大弯	81	372	13	8
免疫	20	101	90	64
伴恶心	10	10	7	1
广泛转移	4	3	2	10
弱阳性	1	5	12	12
…	…	…	…	…

(2) 临床术语–治疗方案矩阵的奇异值分解

奇异值分解(Singular Value Decomposition, SVD)是LSA中构造语义空间的常见方法之一，大量应用于解决不受限的最小立方问题、矩阵阶次估计和规范相关分析等问题。通过矩阵的奇异值分解，可得到矩阵A的三个原始矩阵乘积的形式：

$$A = U\sum V^T \tag{2}$$

其中，U是$m \times r$的A的左奇异正交矩阵，U的列向量称为左奇异值向量。V是$r \times n$的A的右奇异正交矩阵，V的行向量称为右奇异值向量。$\sum$是$r \times r$的A的奇异值组成的对角矩阵。

在本实验中，利用 NumPy[13]实现矩阵的奇异值分解，将临床术语–治疗方案矩阵分解成三个矩阵 U、S、V，其中矩阵 U 的 1 092 个行向量代表 1 092 个临床术语在语义空间中的坐标向量，矩阵 V 的 4 个行向量代表 4 种治疗方案在语义空间中的坐标向量。由于矩阵 U 和矩阵 V 的维度大，计算较为复杂，因此本研究将 U 和 V 分别投影到二维、三维和四维语义空间内，然后计算临床术语和治疗方案在二维、三维和四维语义空间内的坐标向量[14]。

对所得的矩阵进行奇异值分解，根据矩阵 U 和 V，分别取矩阵 U 和 V 的前两列、前三列和全部四列，分别可以得到临床术语和治疗方案在二维、三维和四维语义空间内的坐标向量，如表 4 至表 9 所示：

表 4 临床术语在二维语义空间内的坐标向量示例

临床术语	二维语义空间坐标	
无压痛	–0.1007	–0.0638
大弯	–0.0384	0.0662
免疫	–0.0139	–0.0439
伴恶心	–0.0014	–0.0016
广泛转移	–0.0009	–0.0028
弱阳性	–0.0011	–0.0088

表 5 临床术语在三维语义空间内的坐标向量示例

临床术语	三维语义空间坐标		
无压痛	–0.1007	–0.0638	0.0045
大弯	–0.0384	0.0662	0.0107
免疫	–0.0139	–0.0439	0.0122
伴恶心	–0.0014	–0.0016	–0.0026
广泛转移	–0.0009	–0.0028	–0.0072
弱阳性	–0.0011	–0.0088	–0.0012

表 6　临床术语在四维语义空间内的坐标向量示例

临床术语	四维语义空间坐标			
无压痛	–0.1007	–0.0638	0.0045	0.0945
大弯	–0.0384	0.0662	0.0107	–0.0199
免疫	–0.0139	–0.0439	0.0122	0.0353
伴恶心	–0.0014	–0.0016	–0.0026	0.0001
广泛转移	–0.0009	–0.0028	–0.0072	0.0010
弱阳性	–0.0011	–0.0088	–0.0012	–0.0025

表 7　治疗方案在二维语义空间内的坐标向量

治疗方案	二维语义空间坐标	
手术+放化疗	–0.2311	0.0022
手术	–0.9191	0.3270
对症治疗	–0.2469	–0.7655
放化疗	–0.2023	–0.5542

表 8　治疗方案在三维语义空间内的坐标向量

治疗方案	三维语义空间坐标		
手术+放化疗	–0.2311	0.0022	–0.6432
手术	–0.9191	0.3270	0.1587
对症治疗	–0.2469	–0.7655	0.4815
放化疗	–0.2023	–0.5542	–0.5738

表 9　治疗方案在四维语义空间内的坐标向量

治疗方案	四维语义空间坐标			
手术+放化疗	–0.2311	0.0022	–0.6432	0.8592
手术	–0.9191	0.3270	0.1587	–0.2075
对症治疗	–0.2469	–0.7655	0.4815	0.2722
放化疗	–0.2023	–0.5542	–0.5738	–0.3802

3.5　潜在语义空间的应用

由临床术语–治疗方案矩阵构建的潜在语义空间能充分体现临床术语与临床术语之间、临床术语与治疗方案之间以及治疗方案与治疗方案之间的相关关系。要查询单个或多个临床术语与某一治疗方案的相关关系，只需将所查询的临床术语投影到所构建的潜在语义空间中，计算出其在语义空间内的坐标向量，采用余弦夹角定理，即可计算出临床术语或临床术语组合的坐标向量与各治疗方案之间的语义距离。根据各临床术语的出现频率生成查询向量 Xq[15]，按公式(3)对 Xq 进行处理。

$$Dq = X_q^T U S^{-1} \qquad (3)$$

其中，Dq为待查询向量在语义空间中的坐标向量。

通过余弦夹角公式[16]，可以计算出临床术语或临床术语组合与治疗方案在语义空间内的相关度，公式如下：

$$Cq = \frac{\sum_{k=1}^{n} Dq_i D_i}{\sqrt{\sum_{k=1}^{n} (Dq_i^2)}\sqrt{\sum_{k=1}^{n} (D_i^2)}} \qquad (4)$$

其中，Cq 为查询向量在语义空间内的坐标向量 Dq 与某治疗方案在语义空间内的坐标向量 D 之间的相关度大小，Dq_i 为向量 Dq 的第 i 个分量，D_i 为向量 D 的第 i 个分量，n 为语义空间的维度大小。在 NumPy 中，编写程序计算查询向量与所有治疗方案的相关度，并以相关度值最大的治疗方案作为决策结果，建立治疗方案选择的决策支持模型。

3.6　决策支持模型的评价

对于所建立的决策支持模型，可以通过准确性进行评价，其步骤与以上建模过程相似。对测试样本进行治疗方案抽取、分词处理；经奇异值分解构建相应的二维、三维和四维语义空间；计算出相关度，并根据相关度大小决定每个样本的治疗方案，将计算出的治疗方案与实际的治疗方案进行对照，一致者为准确预测，不一致为错误预测。

例如，对于图 1 的出院记录，从中抽取治疗方案，对其文本进行分词，并依据其出现频次构建查询向量 Xq。将 Xq 投影到语义空间中，分别得到其在二维语义空间的坐标向量 Dq_2(–0.0008, –0.0029)、三维语义空间内的坐标向量 Dq_3(–0.0008, –0.0029, –0.0010)和四维语义空间的坐标向量 Dq_4(–0.0008, –0.0029, –0.0010, 0.0024)。通过余弦公式计算此向量 Xq 与各治疗方案向量在同一个语义空间内的相关度。将治疗方案按照相关度大小降序排序，结果如表 10 至表 12 所示：

表 10　二维空间内治疗方案及其相关度

治疗方案	相关度
对症治疗	0.9995
放化疗	0.9976
手术+放化疗	0.2686
手术	–0.0604

表 11　三维空间内治疗方案及其相关度

治疗方案	相关度
放化疗	0.9033
对症治疗	0.6435
手术+放化疗	0.3925
手术	–0.1086

表 12 四维空间内治疗方案及其相关度

治疗方案	相关度
对症治疗	0.6671
手术+放化疗	0.6606
放化疗	0.3968
手术	–0.2165

由表 10 可知，在二维语义空间内，对于图 1 中的出院记录信息推测出的最佳治疗方案为对症治疗，而病历信息中采取的治疗方案为放化疗，即模型推算错误。但是，由表 11 可知，在三维语义空间内，对于图 1 中的出院记录信息推算出的最佳治疗方案为放化疗，这与实际病历信息中所采取的治疗方案是一致的，即模型推算正确。由表 12 可知，在四维语义空间内，对于图 1 中的出院记录信息推测出的最佳治疗方案为对症治疗，而病历信息中采取的治疗方案为放化疗，即模型推算错误。

使用所收集病历中的 1 000 份测试样本，对建立的决策模型进行测试。在二维语义空间内，有 504 份测试样本经决策模型推算出的治疗方案与实际采取的治疗方案一致，准确率为 50.4%。而在三维语义空间内，有 605 份测试样本经决策模型推算出的治疗方案与实际采取的治疗方案一致，准确率为 60.5%。在四维语义空间内，有 529 份测试样本经决策模型推算出的治疗方案与实际采取的治疗方案一致，准确率为 52.9%。

4 结果及讨论

模型的测试结果显示，通过构建电子病历文本的潜在语义空间，可以从历史病历信息中有效地抽取临床决策支持规则，建立决策支持模型。这证明了潜在语义分析方法在医学文本分析中有较好效果。

通过实验发现，在矩阵降维前，模型的准确率为 52.9%，通过对矩阵降维，计算其二维和三维语义空间，发现在二维语义空间中(见表 7)，对症治疗和放化疗两种治疗方案在语义上较为接近，在计算查询向量与治疗方案的相关度时，两者的相关度大小接近，导致模型推算的准确率较低。而在三维潜在语义空间中，临床术语和治疗方案的坐标向量都较为精确，减少了彼此间的相互干扰，强化了临床术语与治疗方案之间的潜在语义结构，提高了模型推算的准确率(60.5%)。

在潜在语义空间的构建过程中，维数 k 的选择是语义空间构建中的关键。通过查阅文献[17]，本研究中将 k 值分别选择为 2、3 和 4，构建二维语义空间、三维语义空间和四维语义空间。维数越大，临床术语和治疗方案的空间位置越精确，但是会导致 4 种治疗方案的相关度值大小均较低，且数值大小差别不大；维数越小，噪音越大，空间位置不精确，导致治疗方案的相关度值大小均较高。通过实验发现，选择三维语义空间，治疗方案相关度值大小差别大，且空间位置较为精确，能够有效地减少干扰因素，实验准确率相对较高。

5 结 语

本文通过抽取临床术语和治疗方案，构建临床术语–治疗方案矩阵，利用 NumPy 分解矩阵，完成了语义空间的构建、应用和评价。分别构建了二维、三维和四维潜在语义空间。在语义空间内，可以发现临床术语与治疗方案之间的潜在语义结构，从中抽取辅助治疗方案选择的决策知识。研究结果可以证实潜在语义分析技术能够有效地应用到医学文本分析中。

由于存在实验训练样本数量较少，构建的词典不够完备，临床术语的多样性以及治疗方案抽取错误等问题，在三维语义空间内，模型的准确率还有待提高。此外，治疗方案抽取原则的制定至关重要，实验经过大量查阅病历信息以及咨询相关临床医生，并经过反复实验测试，完成了治疗方案抽取原则的制定。虽然抽取得到的治疗方案准确率高，但由于病历信息的多样性，在抽取过程中仍然可能会出现错误。例如，某位患者服用某种特殊的化疗药物，但是其出院记录中并没有出现化疗相关字样，那么，在抽取治疗方案时，系统会认为该患者采用的治疗方案为对症治疗，从而导致治疗方案抽取错误。如何提高治疗方案抽取的准确率，是未来研究内容之一。

参考文献：

[1] Landauer T K. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge [J]. Psychological Review, 1997, 104(2): 211-240.

[2] Cohen T, Blatter B, Patel V. Simulating Expert Clinical

Comprehension: Adapting Latent Semantic Analysis to Accurately Extract Clinical Concepts from Psychiatric Narrative [J]. Journal of Biomedical Informatics, 2008, 41(6): 1070-1087.

[3] Cohen T, Blatter B, Patel V. Exploring Dangerous Neighborhoods: Latent Semantic Analysis and Computing Beyond the Bounds of the Familiar [C]. In: Proceedings of the Annual Symposium of American Medical Informatics Association. 2005: 151-155.

[4] Ginter F, Suominen H, Pyysalo S, et al. Combining Hidden Markov Models and Latent Semantic Analysis for Topic Segmentation and Labeling: Method and Clinical Application [J]. International Journal of Medical Informatics, 2009, 78(12): 1-6.

[5] Wild F, Haley D. Using Latent-Semantic Analysis and Network Analysis for Monitoring Conceptual Development [J]. Journal for Language Technology and Computational Linguistics, 2011, 26(1): 9-21.

[6] Wang J, Sun X P, Nahavandi S, et al. Multichannel Biomedical Time Series Clustering via Hierarchical Probabilistic Latent Semantic Analysis [J]. Computer Methods and Programs in Biomedicine, 2014, 117(2): 238-246.

[7] Abate F, Acquaviva A, Ficarra E, et al. A New Latent Semantic Analysis Based Methodology for Knowledge Extraction from Biomedical Literature and Biological Pathways Databases [C]. In: Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms, Rome, Italy. 2011: 66-74.

[8] 甘艳芳, 倪子伟, 林凡. 潜在语义分析在中医证候分类中的应用[J]. 厦门大学学报: 自然科学版, 2012, 51(6): 991-994. (Gan Yanfang, Ni Ziwei, Lin Fan. The Application of LSA in Traditional Chinese Medicine Syndromes Classification [J]. Journal of Xiamen University: Natural Science, 2012, 51(6): 991-994).

[9] 雷蕾, 张早华, 温先荣, 等. 概率潜在语义分析(PLSA)在中药新药处方发现中的应用[J]. 世界科学技术(中医药现代化), 2012(5): 1976-1980. (Lei Lei, Zhang Zaohua, Wen Xianrong, et al. Study on Application of Probability Latent Semantic Analysis (PLSA) in Herbal Prescription Development [J]. World Science and Technology (Modernization of Traditional Chinese Medicine and Materi Medica), 2012(5): 1976-1980).

[10] 中华人民共和国国家卫生和计划生育委员会. 胃癌规范化诊治指南(试行)[J]. 中国医学前沿杂志(电子版), 2013, 5(8): 29-36. (National Health and Family Planning Commission of the People's Republic of China. Gastric Standardized Treatment Guidelines (Trial) [J]. Chinese Journal of the Frontiers of Medical Science (Electronic Version), 2013, 5(8): 29-36.)

[11] 王思力. 面向大规模信息检索的中文分词技术研究[D]. 北京: 中国科学院研究生院, 2006. (Wang Sili. Research on Chinese Word Segmentation for Large Scale Information Retrieval [D]. Beijing: Graduate School of Chinese Academy of Sciences, 2006.)

[12] Chung Y M, Lee J Y. A Corpus-based Approach to Comparative Evaluation of Statistical Term Association Measure [J]. Journal of the American Society for Information Science and Technology, 2001, 52(4): 283-296.

[13] Idris I. Python 数据分析基础教程 NumPy 学习指南[M]. 张驭宇译. 北京: 人民邮电出版社, 2014: 110. (Idris I. NumPy Beginner's Guide [M]. Translated by Zhang Yuyu. Beijing: Posts & Telecom Press, 2014: 110.)

[14] Bendersky M, Croft W B. Discovering Key Concepts in Verbose Queries [C]. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008: 491-498.

[15] 李国垒, 陈先来. 潜在语义分析在关键词—叙词对照系统构建中的应用[J]. 情报理论与实践, 2014, 37(4): 127-130, 133. (Li Guolei, Chen Xianlai. The Application of Latent Semantic Analysis to Construction of Keyword-Descriptor Comparison System [J]. Information Studies: Theory & Application, 2014, 37(4): 127-130, 133.)

[16] 夏冬, 肖晓旦, 李国垒, 等. 基于潜在语义分析的关键词–分类号对应关系研究[J]. 现代图书情报技术, 2014(12): 92-96. (Xia Dong, Xiao Xiaodan, Li Guolei, et al. Research on Correspondence Between Keyword and Chinese Library Classification Based on Latent Semantic Analysis[J]. New Technology of Library and Information Service, 2014(12): 92-96.)

[17] 盖杰, 王怡, 武港山. 基于潜在语义分析的信息检索[J]. 计算机工程, 2004, 30(2): 58-60. (Gai Jie, Wang Yi, Wu Gangshan. Text Information Retrieval Based on Latent Semantic Analysis [J]. Computer Engineering, 2004, 30(2): 58-60.)

作者贡献声明:

杨荣: 提出研究思路;

陈先来: 设计研究方法, 准备实验环境;

李国垒: 实施实验, 收集数据, 起草论文;

夏冬：论文修改及最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 李国垒. SinoMed.xlsx. 中国生物医学文献数据库术语.
[2] 李国垒. Matrix.csv. 临床术语与治疗方案矩阵.
[3] 李国垒. ClinicalTerms.xlsx. 临床术语.
[4] 李国垒. MutualInformation.xlsx. 互信息值大小.
[5] 李国垒. Semantic Analysis.py. 语义分析代码.

收稿日期: 2015-09-28
收修改稿日期: 2015-11-18

Latent Semantic Analysis of Electronic Medical Record Text for Clinical Decision Making

Li Guolei[1] Chen Xianlai[1,2,3] Xia Dong[4] Yang Rong[5]
[1](Information Security and Big Data Research Institute, Central South University, Changsha 410013, China)
[2](Key Laboratory of Medical Information Research (Central South University), College of Hunan Province, Changsha 410013, China)
[3](Hunan Province Cooperative Innovation Center of Medical Big Data, Changsha 410013, China)
[4](Chengdu Documentation and Information Center, Chinese Academy of Sciences, Chengdu 610041, China)
[5](Xiangya Hospital, Central South University, Changsha 410078, China)

Abstract: [**Objective**] This study aims to extract knowledge for clinical decision from electronic medical records through semantic analysis. [**Methods**] We first extracted clinical terms from the training samples by the word segmentation algorithm with the help of custom dictionary and statistical method. Then, we used latent semantic analysis to find the potential correlations between clinical terms and treatment plans. Finally, we established a latent semantic model to support gastric cancer treatments. [**Results**] We successfully extracted 605 treatment plans from 1000 test samples based on the discharge summary texts. [**Limitations**] Only discharge record texts were examined for this study. [**Conclusions**] The latent semantic analysis could effectively process electronic medical records to assist doctors' clinical decision-making work, which posed positive effects to the development of electronic medical record applications.

Keywords: Electronic medical record Chinese text segmentation Latent Semantic Analysis Gastric cancer Clinical decision support Selection of treatment plans